Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-03-21 16:51:41
REFERENCE NO: 187

Author Names & Affiliations

- Ronald Levy - Professor, Temple University

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Theoretical and Computational Chemistry, Computational Biology and Biophysics

Title of Submission

Molecular Simulations for the Study of Protein Fitness and Free Energy Landscapes, and their Applications in Pharmaceutical Design

Abstract (maximum ~200 words).

Molecular recognition forms the basis for virtually all biological processes. Modern computational tools for modeling molecular recognition at atomic resolution can provide insights not obtainable by experiments alone. At the core of these computer models are effective potential energy functions and algorithms to model conformational dynamics. Our group has an extensive track record for carrying out leading edge research on free energy landscapes of proteins that control molecular recognition, based on state-of-the-art molecular dynamics simulations in structure space, and on protein fitness landscapes constructed using maximum entropy Potts models in sequence space. We are developing and applying these computational tools to design molecules which inhibit the life cycle of the Human Immunodeficiency Virus (HIV), and to target kinase family proteins which are implicated in various cancers. The molecular simulations in structure and sequence space which form the basis of our research require XSEDE scale resources for both tightly and loosely coupled simulations involving thousands to hundreds of thousands of processors.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

I. Protein Free Energy Landscapes in Structure Space
Conformational dynamics plays a fundamental role in the function of biomolecules and statistical mechanics provides the framework for extracting the thermodynamic and kinetic information contained in atomic simulations of these systems. Enhanced sampling techniques are required to map the free energy landscapes, and advanced statistical and computational techniques based on reweighting are required to analyze them. The twin problems of sampling and reweighting are interrelated. Our research program builds on the work of our group over many years to develop and apply innovative computational approaches to map the free energy and fitness landscapes of proteins. This mapping of very high dimensional structure and sequence spaces can be accomplished by constructing biased simulations along collective

variables and then by unbiasing the trajectories during the subsequent analysis phase. There are many ways to do this. Our current research on protein free energy landscapes in structure space focuses on multi-canonical sampling schemes like synchronous replica exchange that are designed for use on tightly coupled high performance XSEDE computer clusters; we have also written asynchronous replica exchange code for use on loosely coupled computational grids which span sizes from thousands of processors (campus grids) to hundreds of thousands of processors (the IBM World Community Grid). Some of our most recent work has focused on mapping protein-ligand binding free energy landscapes using asynchronous replica exchange. We have devised a novel approach to solve the problem of constructing accurate free energy estimates from simulations running on massive but minimally communicating computational grids which involves the following: (1) the development of stochastic reweighting schemes which incorporate knowledge about Markovian States into the stochastic reweighting algorithms in order to correctly account for meta-stable states that are locally but not globally equilibrated on the free energy landscape; (2) the development of adaptive replica exchange algorithms which uses UWHAM weights determined stochastically to assign replicas to thermodynamic states.

Xia, J., W.F. Flynn, E. Gallicchio, B. Zhang, P. He, Z. Tan, and R.M. Levy (2015). Large Scale Asynchronous and Distributed Multi-Dimensional Replica Exchange Molecular Simulations and Efficiency Analysis. Journal of Computational Chemistry, 36 (23), 1772-1785. doi: 10.1002/jcc.23996, PMCID: PMC4512903
.
Zhang, Bin W., Junchao Xia, Zhiqiang Tan, and R.M. Levy (2015). A Stochastic Solution to the Unbinned WHAM Equations. Journal of Physical Chemistry Letters, 6, 3834-3840. DOI: 10.1021/acs.jpclett.5b01771.

Zhang, Bin W., Wei Dai, Emilio Gallicchio, Peng He, Junchao Xia, Zhiqiang Tan, and R.M. Levy (2016). Simulating Replica Exchange: Markov State Models, Proposal Schemes and the Infinite Swapping Limit. Journal of Physical Chemistry B, 20 (33), 8289–8301 DOI: 10.1021/acs.jpcb.6b02015.

II. Protein Fitness Landscapes in Sequence Space
Proteins are constrained structurally, functionally, and thermodynamically. These constraints restrict the set of allowable mutations at each site in a protein's sequence. When examined in a bulk collection of homologous sequences, these constraints manifest as covarying or correlated sets of mutations patterns. We are pioneering new sequence-based statistical inference methods to extract the underlying structural and functional signals from multiple sequence alignments (MSAs) of homologous protein sequences. These methods encode the sequence variation present in an MSA into pairwise spin-glass Hamiltonian models of the sequence space. The pairwise nature of the Hamiltonian models allows the models to capture the context-dependent effects of a mutation, and we are using these models to analyze correlated mutations and the fitness landscapes of kinase family protein allostery and of HIV proteins evolving under drug selection pressure.

These models are based on the maximum entropy principle that seeks to construct the minimally biased sequence probability distribution that reproduces the pair correlations present in an MSA. Due to the similarity to the Ising spin model in statistical physics, the procedure of inferring a maximum likelihood model from data observables has been called the Inverse Ising problem. Solving the Inverse Ising problem is a significant computational challenge because the number of model parameters scales with the square of the protein's length. A variety of techniques have been developed to solve this inverse inference problem, all with various degrees of approximation. The most approximate methods operate on a scale of minutes on a single CPU core, while the least approximate methods can take hundreds of hours across multiple CPU threads. We have developed a GPU-based Markov Chain Monte Carlo (MCMC) inference methodology that leverages the parallel computing power of GPUs to find the optimal set of model parameters through multidimensional Newton search in several hours. While substantially faster than CPU implementations of MCMC-based inverse inferences, the success of our implementation is contingent on the availability of computing hardware that supports multiple modern GPUs.

Haldane, Allan, William F. William, Peng He, R. S. K. Vijayan, and R.M. Levy (2016). Structural propensities of kinase family proteins from a Potts model of residue co-variation. Protein Science, 25, 1378- 1384. DOI: 10.1002/pro.2954

Flynn, William F., Allan Haldane, Bruce E. Torbett, and Ronald M. Levy (2017). Inference of epistatic effects and the development of drug resistance in HIV-1 protease. Molecular Biology and Evolution, DOI: 10.1093/molbev/msx095

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data

infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

We would like to increase our use of very large scale loosely distributed computing resources but the current instrumentation supported by the Office of NSF Cyberinfrastructure is not suitable for many grid computing applications. The only supported instrument of which we're aware that utilizes grid computing resources is the Open Science Grid that hosts a Condor pool of ~80,000 CPU cores. While a step in the right direction, this resource is limited in several ways. Submitted jobs do not support multithreading and the number of cores available at one time to a single user or group is very limited. For comparison, the entire Open Science Grid is only about 10 times larger than resources available on a single modest campus grid.

For work with statistical inferences of protein mutational and fitness landscapes, utilizing GPU processing power is critical for parameterizing accurate models quickly. While several computing resources available through XSEDE support GPU and CPU/GPU computing, there exists only one resource (XStream) that is designed for GPU-intensive computing, though it is limited with only 65 nodes. Moving forward, there should be more NSF office of Cyberinfrastructure supported resources with a GPU/CPU ratio closer to one (XStream boasts 0.8 GPUs/CPU core). Certain computations can be substantially accelerated by increasing the number of GPUs per node as internode communication (e.g. through MPI) imposes serious limitations on memory transfer speeds when compared to PCIe transfers within a node. In addition, this will allow for more efficient use of service units for GPU-only or GPU-heavy computations so that a dozen or more CPU cores do not sit idle while GPU-based computations are underway.

In addition to the absence of certain types of cyberinfrastructure, there are several difficulties working with the currently supported infrastructure. Working across several HPC systems can be challenging, primarily because they each have different environments. Moving a computation from one machine to another can be challenging as it will require recompiling or installing software and modifying or rewriting submission and organization scripts. The cost of recompiling some software packages in different environments can be very high and the level of technical support varies between institutions. A more unified building, storage, and submission interface could greatly alleviate some of these pain points. That said, more detailed information regarding each clusters' infrastructure should be more easily accessible. For example, as far as we know, users are not privy to the underlying network topology at each support computing center; this information can be critical for tightly-coupled computing jobs where a relatively slow network interconnection will slow down the entire computation. Detailed information of this nature should be more easily accessible.
Lastly, the application process for XSEDE resources can be improved. An application must be submitted each year despite the fact that the XSEDE resources are often being used to support multiyear research grants. The reviews can be terse and little attempt is made to reconcile conflicting reviews in the context of resource allocations. Doing reviews less frequently but more thoroughly might help the process.

Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode)."